Leveraging large language models for predictive chemistry

Year: 2024 | Citations: 288 | Authors: K. Jablonka, P. Schwaller, Andres Ortega■Guerrero

Abstract

Machine learning has transformed many fields and has recently found applications in chemistry and materials science. The small datasets commonly found in chemistry sparked the development of sophisticated machine learning approaches that incorporate chemical knowledge for each application and, therefore, require specialized expertise to develop. Here we show that GPT-3, a large language model trained on vast amounts of text extracted from the Internet, can easily be adapted to solve various tasks in chemistry and materials science by fine-tuning it to answer chemical questions in natural language with the correct answer. We compared this approach with dedicated machine learning models for many applications spanning the properties of molecules and materials to the yield of chemical reactions. Surprisingly, our fine-tuned version of GPT-3 can perform comparably to or even outperform conventional machine learning techniques, in particular in the low-data limit. In addition, we can perform inverse design by simply inverting the questions. The ease of use and high performance, especially for small datasets, can impact the fundamental approach to using machine learning in the chemical and material sciences. In addition to a literature search, querying a pre-trained large language model might become a routine way to bootstrap a project by leveraging the collective knowledge encoded in these foundation models, or to provide a baseline for predictive tasks. Machine learning techniques are widely employed in chemical science, but are application specific and their development requires dedicated expertise. Jablonka and colleagues fine-tune the GPT-3 model and show that it can provide surprisingly accurate answers to a wide range of chemical questions.